Inteligencia artificial y evaluación en contextos universitarios: una revisión de los últimos avances

Artificial Intelligence and Assessment in University Contexts: A Review of Recent Advances

Alejandra Alexia Díaz Pino
Doctora en Educación. Profesora titular CES Don Bosco

Resumen

La inteligencia artificial ha traído consigo cambios constantes y dinámicos en el ámbito educativo, especialmente en lo que respecta a la simulación de entornos reales, personalización del aprendizaje, accesibilidad, equidad y optimización de tareas complejas, como es el proceso de evaluación. El objetivo del presente estudio es analizar el impacto de los modelos basados en inteligencia artificial en ámbitos universitarios. Con el ánimo de identificar potencialidades, limitaciones y estrategias esenciales para su integración eficaz, se realiza una revisión sistemática de los últimos avances mediante el Protocolo Prisma. Gran parte de los estudios seleccionados muestran avances sustanciales en cuanto al diseño de instrumentos, calificación y retroalimentación; aunque condicionados a instrucciones contextualizadas, vinculación con elementos curriculares y tipo de material a evaluar y supervisión humana. La formación del docente y estudiante, selección de instrumentos adecuados y desarrollo técnico de esta tecnología, junto su uso ético resulta fundamental para su integración eficaz y responsable.

Palabras clave: inteligencia artificial, evaluación, diseño de instrumentos, calificación, universidad.

Abstract

Artificial intelligence (AI) has brought continuous and dynamic changes to the educational field, particularly in simulating real environments, personalising learning, accessibility, equity, and optimising complex tasks such as assessment processes. This study aims to analyse the impact of AI-based models in university contexts. A systematic review of recent advances was conducted using the PRISMA Protocol to identify potentials, limitations, and essential strategies for effective integration. Most selected studies reveal substantial progress in the design of assessment instruments, grading, and feedback. However, their effectiveness depends heavily on contextualised instructions, alignment with curricular elements, the type of material assessed, and human oversight. Essential components for effective and responsible integration include training for teachers and students, appropriate instrument selection, technical development of the technology, and ethical use.

Keywords: artificial intelligence, assessment, instrument design, grading, university.

ISSN: 1576-5199
Fecha de recepción: 28/01/2025
Fecha de aceptación: 22/03/2025

Educación y Futuro, 52 (2025), 67-83
https://doi.org/10.5281/zenodo.15174010

1. Introducción

El desarrollo de la inteligencia artificial (IA) en los últimos años ha supuesto un avance científico sin precedentes en varios ámbitos como la salud (véase Maslej et al., 2023), la industria o la educación (Farrelly y Baker 2023). En el contexto universitario, ofrece grandes oportunidades para automatizar trabajo repetitivo, simular experiencias de aprendizaje, facilitar metodologías o enfoques activos como el trabajo por proyectos o problemas, así como para promover la inclusión (Flores-Vivar y García-Peñalvo, 2023).

En lo que respecta a la evaluación, los sistemas basados en IA pueden ayudar en el diseño de instrumentos de evaluación complejos a través de instrucciones específicas y claras, automatizar la calificación de grandes cantidades de información y desarrollar análisis e informes exhaustivos ofreciendo información personalizada (Algahtani, 2024). Es fundamental tener en cuenta que los rastros o interacciones con el estudiante permanecen en el sistema facilitando información muy relevante sobre su aprendizaje (Celik, 2023; Ouyang et al., 2023; Zhai y Nehm, 2023).

Además, esta tecnología presenta grandes potencialidades para evaluar de forma equitativita distintas habilidades y niveles. A través de algoritmos, permite analizar el aprendizaje del estudiante de forma personalizada en tiempo real, perimiendo ajustar la prueba a sus necesidades (Alkhuzaey et al., 2024), con las implicaciones que esto tiene para detectar futuros casos de fracaso o abandono escolar y facilitar la intervención educativa (Bañeres et al., 2023).

Asimismo, con los últimos avances de modelos de lenguaje extenso (LLM) como *Gemini*, es posible evaluar imágenes, vídeos o audios (Almasre, 2024; Teckwani et al., 2024), lo cual presenta un avance sustancial que se suma a las opciones que estas herramientas comienzan a ofrecer para la automatización de tareas y el futuro cercano de los asistentes virtuales que se vislumbra con *Operator de OpenAI, Claude AI o Perplexity*. Aunque por restricciones normativas, parte de estas funcionalidades todavía no se encuentran disponibles en Europa (Ley de IA, 2024), ya se han comenzado a utilizar en Estados Unidos.

No obstante, para hacer un buen uso de la IA, es fundamental tener en cuenta las características que debe tener una evaluación que realmente fomente

procesos de enseñanza y aprendizaje de calidad. Así pues, requiere tener un carácter auténtico y evidenciar el nivel competencial del estudiante (Garay-Rondero et al., 2024) e implica evitar centrarse exclusivamente en calificar los aprendizajes, fomentando una evaluación formativa que promueva la formación continua (Álvarez-Herrero, 2020). Todo ello, teniendo en cuenta que el fin último de este proceso siempre debe ser promover la autoevaluación (Luch-Molins y Cano-García, 2022) y facilitar que el estudiante desarrolle competencias metacognitivas basadas en la planificación, supervisión y evaluación de su propio aprendizaje.

La necesidad de abordar las potencialidades que ofrece la inteligencia artificial para facilitar la evaluación justifica realizar estudios cuyo objetivo sea analizar las posibilidades e impacto de la inteligencia artificial para facilitar la evaluación, especialmente en contextos universitarios. La Universidad debe liderar la investigación de la IA en el ámbito educativo promoviendo un uso transparente (Hsu y Ching, 2023) seguro (Chen, 2022; Bahroun et al., 2023), inclusivo (Tlili et al., 2023) y libre de sesgos (Ellaway y Tolsgaard, 2023).

Debido al vertiginoso desarrollo de la IA, comprender este fenómeno emergente requiere que estos estudios se realicen de forma frecuente. Se ha evidenciado un gran paso desde que, hace solo tres años, Gardner et al. (2021) evidenciara que la evaluación con IA aún no era una realidad posible; hasta que revisiones como la de Garay-Rondero et al. (2024) revelaran sus beneficios para la personalización y automatización. Por tanto, parece pertinente realizar estudios, como el presente, cuyo objetivo sea realizar una revisión sistemática basada en el Protocolo Prisma, que analice el impacto de la inteligencia artificial en el ámbito universitario durante los años 2024 y 2025. Todo ello desde una perspectiva pedagógica, diferenciada del enfoque estrictamente computacional que presenta gran parte de la producción científica.

2. Método

La realización de la presente revisión sistemática se basa en la declaración PRISMA 2020 (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*). La declaración PRISMA 2020 reemplaza a la de 2009

e incluye una nueva guía de informes que refleja los avances en los métodos para identificar, seleccionar, evaluar y sintetizar estudios (Page, 2021).

De forma concreta, esta revisión se enfoca en responder ¿qué impacto tiene el uso de la inteligencia artificial en el aprendizaje en contextos universitarios? El objetivo general de esta revisión es, por tanto, analizar el impacto de la IA en la evaluación del aprendizaje en el ámbito universitario, para lo que se establecen los siguientes objetivos específicos:

- Identificar qué beneficios tiene el uso de la IA en la evaluación de la Educación Superior.
- Identificar las limitaciones que tiene el uso de la IA en la evaluación de la Educación Superior.
- Analizar aspectos esenciales para integrar eficazmente la IA en la evaluación universitaria.

Tal y como puede observarse en la *tabla 1*, los criterios de inclusión contemplan todos los estudios empíricos indexados en las bases de datos *Web of Science* y *Scopus* y publicados entre el año 2024 y 2025 cuyo enfoque sea de interés pedagógico y se enmarquen en el ámbito universitario.

Tabla 1
Criterios de inclusión y exclusión

Criterios de inclusión	Criterios de exclusión
Publicados entre 2024 y 2025	Publicados antes de 2024
Indexadas a *Web of Science y Scopus*	No indexadas en *Web of Science o Scopus*
Enfoque pedagógico	Otros enfoques
Ámbito universitario o acceso	Otras etapas educativas
Estudios empíricos	Estudios teóricos, revisiones, metaanálisis

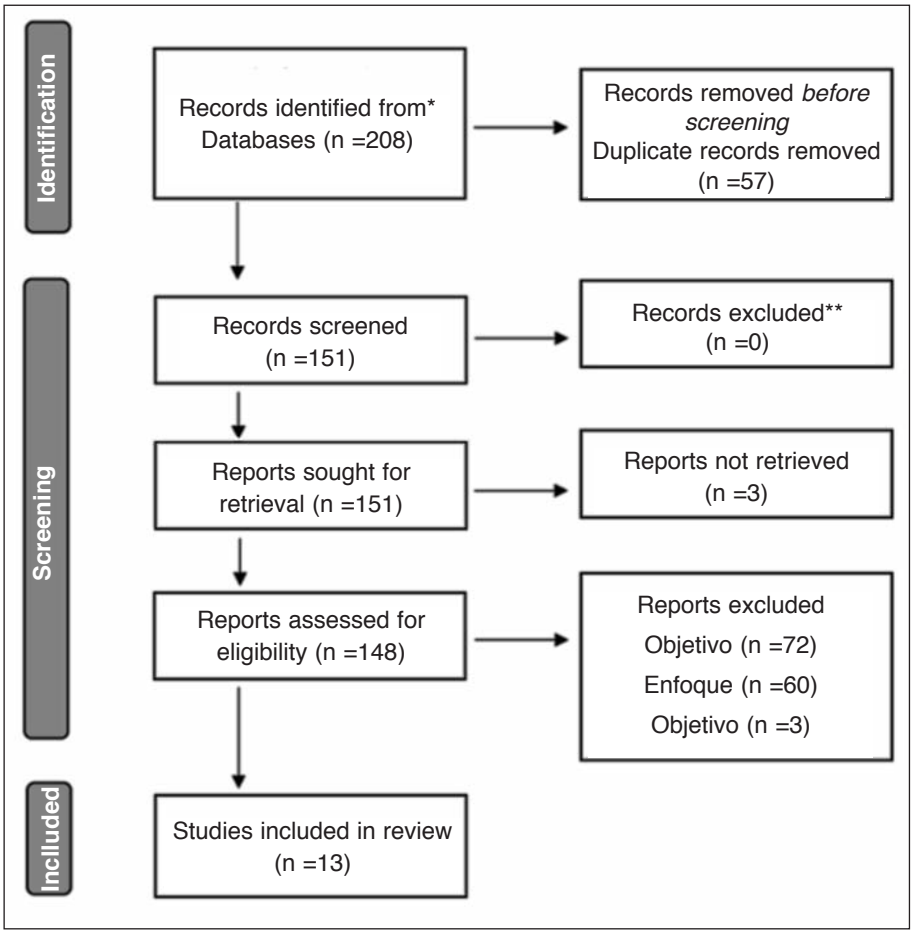
Para la identificación de artículos que analicen las dos variables del estudio: inteligencia artificial y evaluación en entornos universitarios, se utiliza la siguiente ecuación de búsqueda: *(«artificial intelligence» OR «AI»*

OR «machine learning» OR «deep learning») AND («learning assessment» OR «student evaluation» OR «educational assessment»).

El proceso de selección realizado a través de un diagrama de flujo e incluyó inicialmente 208 estudios, de los cuales 77 pertenecen a la base de datos *Web of Science* y 131 a *Scopus*. De la suma inicial de estudios, 57 son descartados por encontrarse duplicados y 3 por imposibilidad de recuperación.

A continuación, 77 estudios fueron excluidos por contar con objetivos diferentes, 60 por partir de un enfoque estrictamente computacional y 3 por tener un objetivo diferente. Finalmente, el número de estudios incluidos ascendió a 14 (véase *figura 1*).

Figura 1
Diagrama de flujo



La distribución geográfica de los estudios seleccionados (véase *tabla 2*) es heterogénea y las áreas de investigación se diferencian en: pedagogía, medicina, diseño, programación, marketing y aprendizaje de idiomas. De los 13 estudios incluidos, 8 se centran en comparar la eficacia de modelos basados en IA y calificadores humanos. De forma concreta (Kyung, 2024; Ruiz et al., 2024; Hamzaoui et al., 2024) se enfocan en el diseño de *ítems;* mientras que (Almasre, 2024; Karademir y Alper, 2024; Kooli y Yusuf, 2024; Teckwani et al., 2024; Mohamed et al., 2024) estudian la calificación.

La investigación de Fernández-Sánchez et al., 2025 se enmarca en analizar las potencialidades de esta tecnología para el diseño de rúbricas, mientras que (Mendonça, 2024; Gritti et al., 2024) analizan el desempeño de la IA en función del material a evaluar y el instrumento; Furze et al. (2024) valora el impacto de la Escala AIAS en el desarrollo de evaluaciones éticas y Cowan et al. (2024) estudia el impacto de la evaluación asistida por IA en la percepción de competencia docente en función del género.

Tabla 2
Estudios seleccionados

Nº	Autor/año	País	Título	Objetivo	Resultados
1	Fernández-Sánchez et al.	España	*Navigating the Future of Pedagogy: The Integration of AI Tools in Developing Educational Assessment Rubrics.*	Analizar el uso de la IA en el diseño de rúbricas.	Agiliza el proceso, mayor precisión y eficiencia en términos de tiempo, garantiza la inclusión de elementos curriculares esenciales.
2	Karademir & Alper, 2024	Turquía	*Evaluating the evaluators: A comparative study of AI and teacher assessments in higher education.*	Comparar calificaciones otorgadas por IA y docentes en función del formato del examen.	Consistencia alta en exámenes tipo test, media cuando se basan en imágenes y baja en vídeos, así como en exámenes tradicionales.
3	Kooli & Yusuf, 2024	No especificado	*Transforming Educational Assessment: Insights into the Use of ChatGPT and Large Language Models in Grading.*	Comparar calificaciones otorgadas por ChatGPT y docentes.	Correlaciones estadísticamente significativas, pero no particularmente sólidas.

Nº	Autor/año	País	Título	Objetivo	Resultados
4	Kyung, 2024	Corea	*A comparative study of AI-human-made and human-made test forms for a university TESOL theory course.*	Comparar la eficacia de ChatGPT y diseñadores humanos en la creación de *ítems.*	Desempeño comparable entre *ítems* creados por IA y humanos.
5	Ruiz et al., 2024	Brasil	*Criação e julgamento de itens: ChatGPT como designer e juiz.*	Comparar la eficacia de ChatGPT y diseñadores humanos en la creación de *ítems*.	Desempeño comparable entre *ítems* creados por IA y humanos.
6	Teckwani et al., 2024	Singapur	*Accuracy and reliability of large language models in assessing learning outcomes achievement across cognitive domains.*	Comparar la eficacia en la calificación de modelos LMMs y diseñadores humanos.	En su mayoría concordancia moderada o baja.
7	Almasre, 2024	Arabia Saudita	*Development and Evaluation of a Custom GPT for the Assessment of Students' Designs in a Typography Course.*	Comparar el desempeño de un GPT personalizado y humano para calificar y brindar retroalimentación.	En su mayoría desempeño similar.
8	Mendonça, 2024	Brasil	*Evaluating ChatGPT-4 Vision on Brazil's National Undergraduate Computer Science Exam.*	Investiga el desempeño de ChatGPT-4 Visión en la resolución de exámenes.	ChatGPT-4 Visión muestra potencialidades prometedoras en evaluaciones académicas multimodales, pero continúa siendo vital la supervisión humana.
9	Mohamed et al., 2024	Sin especificar	*Hands-on analysis of using large language models for the auto evaluation of programming assignments.*	Explorar el potencial de los LLM para automatizar la evaluación de tareas de programación en comparación con evaluadores humanos.	Desempeño comparable a evaluadores humanos en (Gemini y GPT-4o). Mientras que Llama 370B y Mixtral 87B muestran una baja precisión y alineación, particularmente en la resolución de problemas.

Nº	Autor/año	País	Título	Objetivo	Resultados
10	Hamzaoui et al., 2024	Varios	*A hybrid approach for assessing distance learning instructional videos using artificial intelligence.*	Comparar el uso de la IA para la generación automática de preguntas y elaboradas por humanos a partir de las transcripciones de vídeos educativos.	Desempeño comparable, en algunos casos superior de la IA en términos de precisión, relevancia y capacidad para diversificar las estrategias de evaluación.
11	Cowan et al., 2024	Reino Unido y Estados Unidos	*Can AI Level the Playing Field? How AI-Assisted Assessment Impacts Gender Bias in Student Evaluations of Marketing Instructors.*	Analizar el impacto de la evaluación asistida por IA en la percepción de competencia de los docentes en función del género.	Se reducen sesgos en función de género, pero la percepción competencial del docente general también disminuye con el uso de la IA en la evaluación.
12	Furze et al., 2024	Vietnam	*The AI Assessment Scale (AIAS) in action: a pilot implementation of GENAI supported assessment.*	Explorar el impacto de la escala AIAS para promover un uso ético de la IA en la evaluación educativa.	Reducción de los casos de conducta inapropiada en el uso de la IA en la evaluación.
13	Gritti et al., 2024	Canadá	*Progression of an Artificial Intelligence Chatbot (ChatGPT) for Pediatric Cardiology Educational Knowledge Assessment.*	Comparar el rendimiento de ChatGPT 3.5 y 4.0 en evaluaciones educativas.	Desempeño superior en precisión de ChatGPT 4.0 con respecto a la versión 3.5.

3. Resultados

3.1 Identificar qué beneficios tiene el uso de la IA en la evaluación de la Educación Superior

Entre las principales posibilidades que se identifican para facilitar la evaluación a través de la inteligencia artificial se encuentran los siguientes:

3.1.1 Construcción de instrumentos de evaluación

Este tipo de modelos generan *ítems* a gran escala de gran calidad (Hamzaoui et al., 2024), de un modo ágil, que incluye diferentes niveles de dificultad en función de las necesidades, con el tiempo liberado que esto supone para el docente (Kyung, 2024). Asimismo, el diseño de rúbricas se ve particularmente beneficiado por el uso de esta tecnología, ya que facilita su contextualización y alineación con los objetivos de aprendizaje, conjugando el adecuado equilibrio entre especificidad y flexibilidad que implica evaluar partiendo de los elementos prescriptivos del currículo (Fernández-Sánchez et al., 2025).

3.1.2 Calificación automatizada

El uso de esta tecnología presenta grandes potencialidades para la calificación automatizada, estudios como el de Mohamed et al. (2024) indican que los modelos LLM tienen la capacidad de ofrecer una calificación precisa y consistente. Especialmente en los últimos meses, puesto que modelos como *Gemini 1.5 Pro* y *GPT-4o*, han presentado avances significativos para una calificación automatizada confiable. Esos modelos, además, pueden calificar de una forma más objetiva, facilitando evaluaciones imparciales que mitiguen los sesgos del propio docente y arbitren en caso de desacuerdo con el estudiante (Kooli & Yusuf, 2024).

3.1.3 Retroalimentación personalizada

Del mismo modo, este tipo de modelos también pueden ser beneficiosos para proporcionar retroalimentación personalizada, ajustada y uniforme. Autores como (Almasre, 2024) muestran, por ejemplo, el modo en que estas herramientas utilizan una terminología más diversa para dar *feedback* en comparación con los evaluadores humanos. La capacidad de la IA para generar comentarios inmediatos, detallados y personalizados, puede aliviar la presión sobre el estudiante y favorecer la mejora (Kooli & Yusuf, 2024).

3.2 Identificar qué limitaciones tiene el uso de la IA en la evaluación de la Educación Superior

A continuación, se describen los principales aspectos identificados en el uso de la IA.

3.2.1 Falta de consenso en la literatura científica

Mientras que los resultados de estudios como el de Mohamed et al. (2024) muestran que el uso de esta tecnología en Educación Superior presenta evaluaciones precisas y confiables, con niveles comparables a la evaluación humana (Almasre, 2024), así como más auténticas y relevantes para contextos reales; investigaciones como la de Teckwani et al. (2024) encuentran una correlación moderada-baja entre ambos, lo cual sugiere la pertinencia de combinar la labor de ambos y necesidad de un mayor avance técnico en estos instrumentos.

3.2.2 Dependencia del tipo de prueba e instrucciones contextualizas y precisas

La calidad de la evaluación impulsada por IA depende, en gran medida, de la precisión y contextualización de la información y recursos con se alimenta esta tecnología (Teckwani et al., 2024), también en lo que respecta a posibles sesgos o discriminación. Asimismo, varía en función del tipo de evaluación y formato, se ha encontrado una consistencia alta cuando se trata de exámenes objetivos tipo verdadero y falso (Kyung, 2024), media cuando se trata de evaluar imágenes y baja en vídeos y exámenes tradicionales Karademir y Alper, 2024).

3.3 Analizar aspectos esenciales para integrar eficazmente la IA en la evaluación universitaria

A continuación, se presentan los aspectos más relevantes para hacer un uso eficaz de la IA en la evaluación universitaria:

3.3.1 Formación docente y estudiante

Los hallazgos sugieren el potencial de la IA para complementar la experiencia humana en la evaluación educativa junto con la necesidad de un aprendizaje adaptativo por parte de los educadores (Teckwani et al., 2024). El uso eficaz de la IA para la evaluación conlleva una formación docente acorde que lo posibilite, tanto en lo que respecta a las posibilidades del uso aplicado de las misma, como en lo que se refiere a reconocer y mitigar usos inadecuados o excesivos en su utilización (Fernández-Sánchez et al., 2025). Asimismo, de acuerdo con el estudio de Cowan et al. (2024) deter-

minados estudiantes se mantienen escépticos ante los beneficios de esta tecnología y su uso puede afectar en la confianza que tienen sobre la capacidad docente; la formación y justificación pedagógica de su uso podría ayudar a minimizar estos sesgos.

3.3.2 Selección de herramientas basada en evidencias

Estudios como el de Gritti et al. (2024) muestra como ChatGPT 4.0 supera significativamente a la versión 3.5 en evaluaciones educativas mostrando una precisión del 66% frente al 38%. El entramiento de las mismas también es fundamental; los resultados apuntan a que el conocimiento del ámbito de enseñanza y aprendizaje es directamente proporcional a los resultados, por lo que se debe promover que el evaluador pueda entrenar su propio GPT personalizado (Almasre, 2024).

3.3.3 Instrucciones claras, detalladas y contextualizadas

Para lograr resultados óptimos en el diseño de instrumentos de evaluación, las instrucciones deben incluir un volumen amplio y específico de información clara y contextualizada, así como conectar las actividades realizadas en clase y los elementos curriculares (Fernández-Sánchez et al., 2025). De la misma forma, el primer estudio de Hamzaoui et al. (2024) pone de manifiesto la importancia de complementar todo ello con herramientas cuyo objetivo sea fomentar la integridad académica, como la escala AIAS (Perkins et al., 2024).

3.3.4 Supervisión humana

A pesar del vertiginoso avance de la inteligencia artificial, la supervisión humana sigue siendo necesario para garantizar la calidad en los procesos de evaluación (Fernández-Sánchez et al., 2025; Mendonça, 2024). Por ejemplo, en la actualidad, la naturaleza de estos modelos no permite reconocer adecuadamente la expresión creativa o la originalidad o necesidades específicas y, como ha pedido observarse, la evaluación de materiales multimodales basados en imágenes, vídeos o audios presentan limitaciones (Kooli y Yusuf, 2024).

4. Discusión y conclusiones

El presente estudio ha alcanzado los objetivos propuestos, proporcionando un análisis específico de los últimos avances de la inteligencia artificial en la evaluación dentro de contextos universitarios, concretamente en lo que respecta a sus beneficios, limitaciones y estrategias para su integración eficaz. Entre los resultados destacan las potencialidades de estos modelos para diseñar, de forma ágil, instrumentos que permitan evaluaciones contextualizadas, basadas en problemas y conectadas con los elementos prescriptivos del currículo. Así como calificaciones más ricas, objetivas y consistentes, a la par que ofrecen retroalimentación automática, personalizada y ajustada a las necesidades del docente y estudiante.

Sin embargo, es importante tener en cuenta que todavía no hay consenso en la literatura científica sobre su impacto real y su uso eficaz parece estar mediado por el instrumento, material a evaluar y, especialmente, la calidad de las instrucciones y formación docente, siendo esa última una condición imprescindible para obtener el máximo partido de esta tecnología. Para ello, resulta especialmente relevante el Marco de Competencias en IA para docentes de la UNESCO (Fengchun, 2024), puesto que puede servir de hoja de ruta para capacitar al docente en el uso de la IA de forma ética y eficaz, minimizando riesgos potenciales y sacando el máximo partido de sus posibilidades.

Asimismo, parece necesario trabajar más en el desarrollo técnico de esta tecnología, la posibilidad de evaluar distintos materiales como audios, imágenes o vídeos es un gran paso para optimizar el proceso, pero se trata de funcionalidades muy novedosas y, de acuerdo con el escaso número de investigaciones que se han centrado en su estudio, debe perfeccionarse, especialmente en lo que respecta a la evaluación de vídeos.

Tal y como ha podido observarse, los resultados coinciden con investigaciones previas como la Garay-Rondero et al. (2024) sobre las ventajas de esta tecnología para personalizar y automatizar parte del proceso y confirman un avance significativo desde que la revisión de Gardner et al. (2021) cuestionaban la viabilidad de implementar esta tecnología. Asimismo, coincide con ambos estudios en poner de manifiesto la necesidad de seguir desarrollando estas herramientas desde un punto de vista técnico.

En cuanto a la vinculación de estos resultados con una evaluación de calidad, los hallazgos de estudios como el de Fernández-Sánchez et al. (2025) o Kyung (2024) muestran las posibilidades que herramientas como Gemini presentan para promover una evaluación auténtica. De forma concreta, posibilitan el diseño de *ítems* alineados con competencias específicas y la construcción de instrumentos que simulen conexos reales a través de un aprendizaje contextualizado y basado en problemas reales.

Estas herramientas también muestran grandes oportunidades para optimizar la evaluación sumativa, ya que tienen la capacidad de analizar en profundidad amplias cantidades de datos de forma objetiva, facilitando la transparencia (Kooli y Yusuf, 2024) y evitando los sesgos que se pueden gestar a través de la relación entre docente y estudiante o mediando entre cualquier desacuerdo sobre la calificación.

Así como para la evaluación formativa mediante f*eedback* personalizado y ajustado en tiempo real, ofreciendo oportunidades de mejora; para lo que resulta especialmente prometedor el diseño de GPT personalizados y centrados en un dominio específico (Almasre, 2024). Esto último, además, promueve el desarrollo de la autoevaluación por parte del estudiante, haciendo posible la experimentación sin miedo al juicio o fracaso y la reflexión sobre el propio aprendizaje.

No obstante, es fundamental tener en cuenta que el presente estudio cuenta con tres limitaciones principales, el reducido número de artículos incluidos, su concentración en investigaciones correspondientes a los años 2024 y 2025, y en áreas específicas como la pedagogía medicina, diseño, programación, *marketing* y aprendizaje de idiomas, lo que limita la generalización de los resultados.

Entre las principales líneas de investigación futura, se propone el fomento de una mayor investigación enfocada a mejorar y analizar el impacto del uso de la inteligencia artificial en la evaluación de materiales multimodales que incluyen imágenes, vídeos y audios., así como en promover una evaluación formativa que fomente el sentido crítico y el aprendizaje autorregulado.

Todo ello teniendo en cuenta que conlleva una formación que posibilite su uso ético y responsable, pues abre la puerta a dilemas complejos relaciona-

dos con la equidad, seguridad y privacidad de datos. Asimismo, parece importante seguir investigando formas de potenciar la colaboración entre este tipo de modelos y evaluadores humanos para crear procesos de evaluación cada vez más ricos, ágiles y seguros.

Finalmente, es importante enfatizar que el uso de inteligencia artificial en la evaluación de contextos universitarios siempre debe guiarse por criterios pedagógicos que se sustenten evidencias, así como priorizar el bienestar del estudiante. Asimismo, la implicación institucional, el diseño de políticas claras y decididas, además de la formación de estudiantes, docentes, equipos y familias es condición fundamental para que el uso de la tecnología transforme positivamente los procesos de enseñanza y aprendizaje.

Referencias bibliográficas

Algahtani, A. (2024). A Comparative Study of AI-Based Educational Tools: Evaluating User Interface Experience and Educational Impact. *Journal of Theoretical and Applied Information Technology*, *102*(5), 1756-1758. http://bit.ly/4juetRM

Alkhuzaey, S., Grasso, F., Payne, T.R., y Tamma, V. (2024). Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches. *Int J Artif Intell Educ,* (34), 862–914. https://doi.org/10.1007/s40593-023-00362-1

Almasre, M. (2024). Development and Evaluation of a Custom GPT for the Assessment of Students' Designs in a Typography Course. *Education Sciences, 14*(2), 148. https://doi.org/10.3390/educsci14020148

Álvarez-Herrero, J. F. (2020). El poder de la anticipación en la evaluación: simulacros de examen y rúbricas en la educación superior. *Aloma: revista de psicologia, ciències de l'educació i de l'esport Blanquerna*, *38*(2), 51-558. https://bit.ly/4hxt0u2

Bahroun, Z., Anane, C., Ahmed, V., y Zacca, A. (2023). Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. *Sustainability, 15*(17), 12983. https://doi.org/10.3390/su151712983

Bañeres, D., Rodríguez-González, M. E., y Guerrero-Roldán, A. E. (2023). An early warning system to identify and intervene online dropout learners [Un sistema de alerta temprana para identificar e intervenir a los alumnos que abandonan la escuela en línea]. *International Journal of Educational Technology in Higher Education, 20*(3). 5 https://doi.org/10.1186/s41239-022-00371-5

Castonguay, A., Farthing, P., Davies, S., Vogelsang, L., Kleib, M., Risling, T., y Green, N. (2023). Revolutionizing nursing education through Ai integration: A reflection on the disruptive impact of ChatGPT. *Nurse Education Today, 129*, 10591. https://doi.org/10.1016/j.nedt.2023.105916

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior*, *138*, 107468.

Chen, Z. (2022). Artificial intelligence-virtual trainer: Innovative didactics aimed at personalized training needs. *Journal of the Knowledge Economy*, *14*, 1-19. https://doi.org/10.1007/s13132-022-00985-0

Cooper, G. (2023). Examining science education in CHATGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*(3), 444-452. https://doi.org/10.1007/s10956-023-10039-y

Cowan, M., Fox, G., y Larson, K. (2024). Can AI Level the Playing Field? How AI-Assisted Assessment Impacts Gender Bias in Student Evaluations of Marketing Instructors. *Journal of Marketing Education*. https://doi.org/10.1177/02734753241303742

Farrelly T, y Baker N. (2023). Generative Artificial Intelligence: Implications and Considerations for Higher Education Practice. *Educ. Sci, 3*(11), 1-14. https://doi.org/10.3390/educsci13111109

Flores-Vivar, J., y García-Peñalvo, F. (2023). Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4). *Comunicar, 74*, 37-47. https://doi.org/10.3916/C74-2023-03

Ellaway, RH., y Tolsgaard, M. (2023). Artificial scholarship: LLMs in health professions education research. *Advances in Health Sciences Education, 28*, 659–664. https://doi.org/10.1007/s10459-023-10257-4

Fernández-Sánchez, A., Lorenzo-Castiñeiras, J. J., y Sánchez-Bello, A. (2025). Navigating the Future of Pedagogy: The Integration of AI Tools in Developing Educational Assessment Rubrics. *European Journal of Education, 60*(1). https://doi.org/10.1111/ejed.12826

Furze, L., Perkins, M., Roe, J., y MacVaugh, J. (2024). The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI-supported assessment. *Australasian Journal of Educational Technology, 40*(4), 38–55. https://doi.org/10.14742/ajet.9434

Garay-Rondero, C. L., Castillo-Paz, A., Gijón-Rivera, C., Domínguez-Ramírez, G., Rosales-Torres, C., y Oliart-Ros, A. (2024). Competency-based assessment tools for engineering higher education: a case study on complex problem-solving. *Cogent Education, 11*(1). https://doi.org/10.1080/2331186X.2024.2392424

Gardner, J., O'Leary, M., y Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning, 37*(5), 1207-1216. https://doi.org/10.1111/jcal.12577

Gritti, M. N., AlTurki, H., Farid, P., & Morgan, C. T. (2024). Progression of an Artificial Intelligence Chatbot (ChatGPT) for Pediatric Cardiology Educational Knowledge Assessment: Knowledge Assessment: Considerable Gains in a Short Time. *Pediatric Cardiology, 45*(2), 309-313. https://doi.org/10.21203/rs.3.rs-3360192/v1

Hamzaoui, R., Bachiri, Y. A., Ouassam, E., Mouncif, H., y Bouikhalene, B. (2024). A hybrid approach for assessing distance learning instructional videos using artificial intelligence. *International Journal on Technical and Physical Problems of Engineering, 16*(3), 234-243. https://bit.ly/3E7XbtF

Hsu, Y. C., y Ching, Y. H. (2023). Generative Artificial Intelligence in Education, Part One: The Dynamic Frontier. *TechTrends, 1*(5), 603-607. https://doi.org/10.1007/s11528-023-00863-9

Karademir, T., y Alper, A. (2024). Evaluating the evaluators: A comparative study of AI and teacher assessments in higher education. *Digital Education Review,* (45). 124-140. https://doi.org/10.1344/der.2024.45.124-140

Kooli, C., y Yusuf, N. (2024). Transforming Educational Assessment: Insights Into the Use of ChatGPT and Large Language Models in Grading. *International Journal of Human–Computer Interaction*, *41*(5), 3388–3399. https:// doi.org/10.1080/10447318.2024.2338330

Kyung, M. O. (2024). A comparative study of AI-human-made and human-made test forms for a university TESOL theory course. *Lang Test Asia, 14*(19). https://doi.org/10.1186/s40468-024-00291-3

Lluch-Molins, L., y Cano-García, E. (2022). Diseño y validación de un instrumento para la evaluación competencial en el marco del desarrollo profesional docente. *Revista Electrónica Interuniversitaria de Formación del Profesorado, 25*(2), 79-91. https://doi.org/10.6018/reifop.514961

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., y Perrault, R. (2023). *Artificial intelligence index report*. Institute for Human-Centered AI, Stanford University.

Mendonça, N. C. (2024). Evaluating ChatGPT-4 Vision on Brazil's National Undergraduate Computer Science Exam. *ACM Transactions on Computing Education, 24*(3), 1-56. https://doi.org/10.1145/3674149

Ruiz, K. K.., Zúñiga., Pedroza, L.H., y López, A. Y. (2024). Criação e julgamento de itens: ChatGPT como designer e juiz. *Texto Livre, 17*. https://doi.org/10.1590/1983-3652.2024.51222

Parlamento Europeo y del Consejo. (2024). *Reglamento (UE) 2024/1689 del, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial*. https://bit.ly/3WBj3E2

Tlili, A., Shehata, B., y Adarkwah, M. A. (2023). Y si el diablo es mi ángel de la guarda: ChatGPT como caso de estudio del uso de chatbots en la educación. Aprendizaje inteligente. *Reinar, 10*(15), 1-24. https://doi.org/10.1186/s40561-023-00237-x

Mohamed, K., Yousef, M., Medhat, W., Mohamed, E. H., Khoriba, G., y Arafa, T. (2024). Hands-on analysis of using large language models for the auto evaluation of programming assignments. *Information Systems, 128,* 102473.

Ouyang, F., Wu, M., y Zheng, L. (2023) Integración de predicción de rendimiento de inteligencia artificial y análisis de aprendizaje para mejorar el aprendizaje de los estudiantes en el curso de ingeniería en línea. *International Journal of Educational Technology in Higher Education*, *20*(4), 1-23. https://doi.org/10.1186/s41239-022-00372-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., y Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Research methods and reporting, 372,* 1-9. https://doi.org/10.1136/bmj.n71

Perkins, M., Furze, L., Roe, J., y MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment. *Journal of University Teaching and Learning Practice, 21*(6). https://doi.org/10.53761/q3azde36

Teckwani, S. H., Wong, A. H., Luke, N. V., y Low, I. C. (2024). Accuracy and reliability of large language models in assessing learning outcomes achievement across cognitive domains. *Advances in Physiology Education, 48*(4), 904-914. https://doi.org/10.1152/advan.00137.2024

Fengchun. M., y Mutlu, C. (2024). *Marco de competencias en IA para docentes*. UNESCO. https://bit.ly/4ha1G5u

Zhai, X., y Nehm, R. H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching, 60*(6), 1390–1398. https://doi-org.bibliotecauned.idm.oclc.org/10.1002/tea.2188

Cita de este artículo (APA, 7ª ed.):

Díaz, A. A. (2025). Inteligencia artificial y evaluación en contextos universitarios: una revisión de los últimos avances. *Educación y Futuro: Revista de Iinvestigación Aplicada y Experiencias Educativas*, (52), 67-83. https://doi.org/10.5281/zenodo.15174010